Benchmarking DAOS APIs on Google Cloud

DAOS Foundation meeting, ISC'24, 2024-05-13

Nicolau Manubens, Adrian Jackson

nicolau.manubens@ecmwf.int
a.jackson@epcc.ed.ac.uk

Google Cloud's Parallelstore

- DAOS on DRAM + NVMe – 6 TB per node
  - Not currently setup for long term storage
- Ideal: 3 GiB/s write and 6 GiB/s read per server node
  - Based on the storage devices used per node
- Default:
  - POSIX DAOS containers mounted with DFUSE + interception,
  - Protection through erasure coding: EC2+1.
- Supports other DAOS APIs as well as disabling redundancy

Benchmarking

- Looking to investigate
  - Overall performance
    - Scalability
    - WAL functionality as well as just standard performance
  - Interface choice
  - Best number of processes to use
  - Client/server node performance ratios
  - Impact of Sharding, Redundancy, Erasure Coding
  - If we get time, comparison with other software on the same/similar hardware (Ceph, Lustre, etc…)

Auto-scaling Slurm client cluster

- Efficient use of client instances to reduce costs

- Hpc-toolkit
  - https://cloud.google.com/hpc-toolkit/docs/overview
  - Automatic deployments of configurable ”blueprints” on Google Cloud

- Slurm cluster blueprint
  - schedmd-slurm-gcp-v6-controller
  - schedmd-slurm-gcp-v6-nodeset
  - Fine-tuned blueprints and node images to support high-performance DAOS I/O

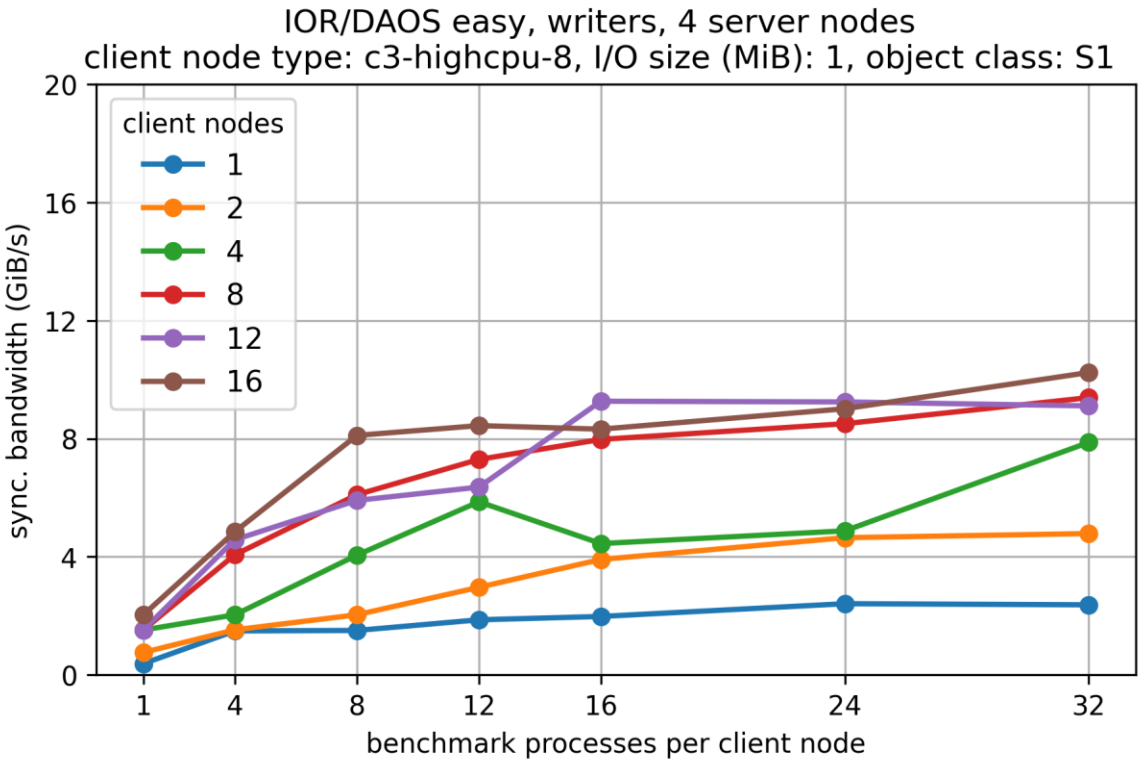
IOR performance of various DAOS APIs

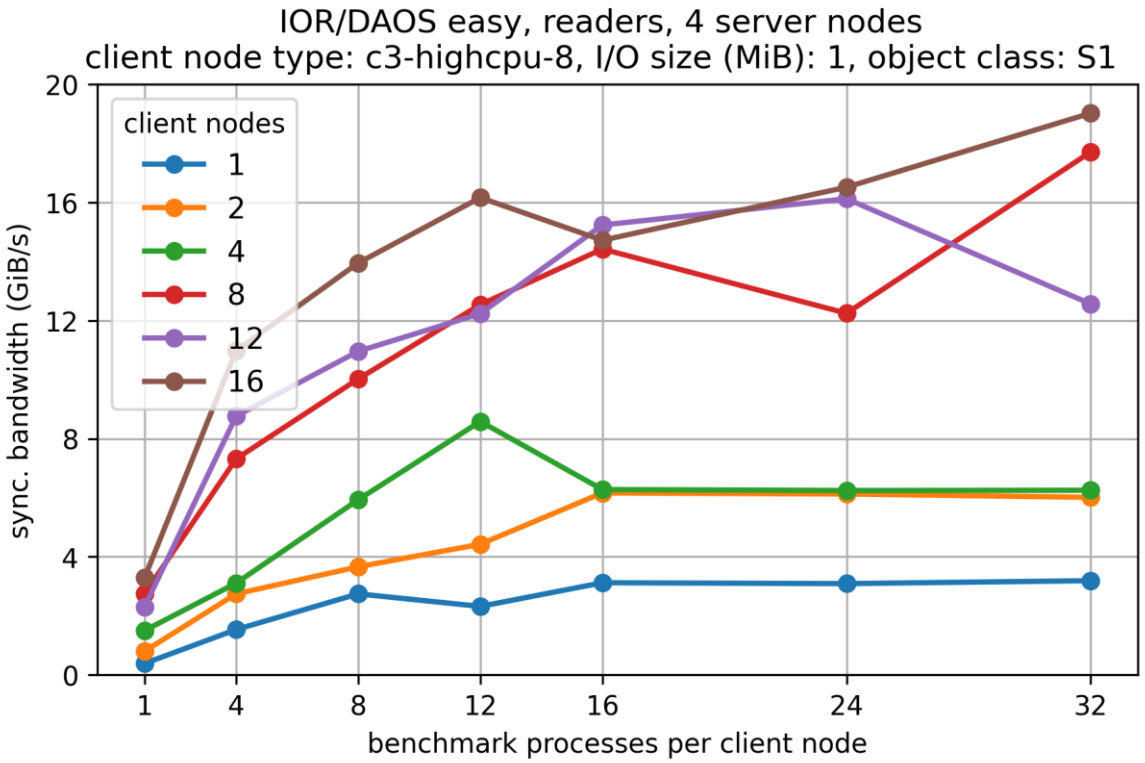
- Redundancy disabled
- 2 client nodes using 4 server nodes
- Initial benchmarks, not particularly tuned for best performance

	IOR bandwidth (GiB/s), 2x c3-highcpu-4, 8 ppn, S1							
	DAOS*	DFS old*	DFS	DFUSE	DFUSE +IOIL	MPIIO+ DFUSE	HDF5+ DFUSE	HDF5 VOL
write	2.0	2.0	2.8	4.5	2-4.6	2.1	3.2	1.6
read	3.2	3.0	3.1	4.7	3.6-5.5	3.4	2.7	err

IOR/DAOS on c3 instances

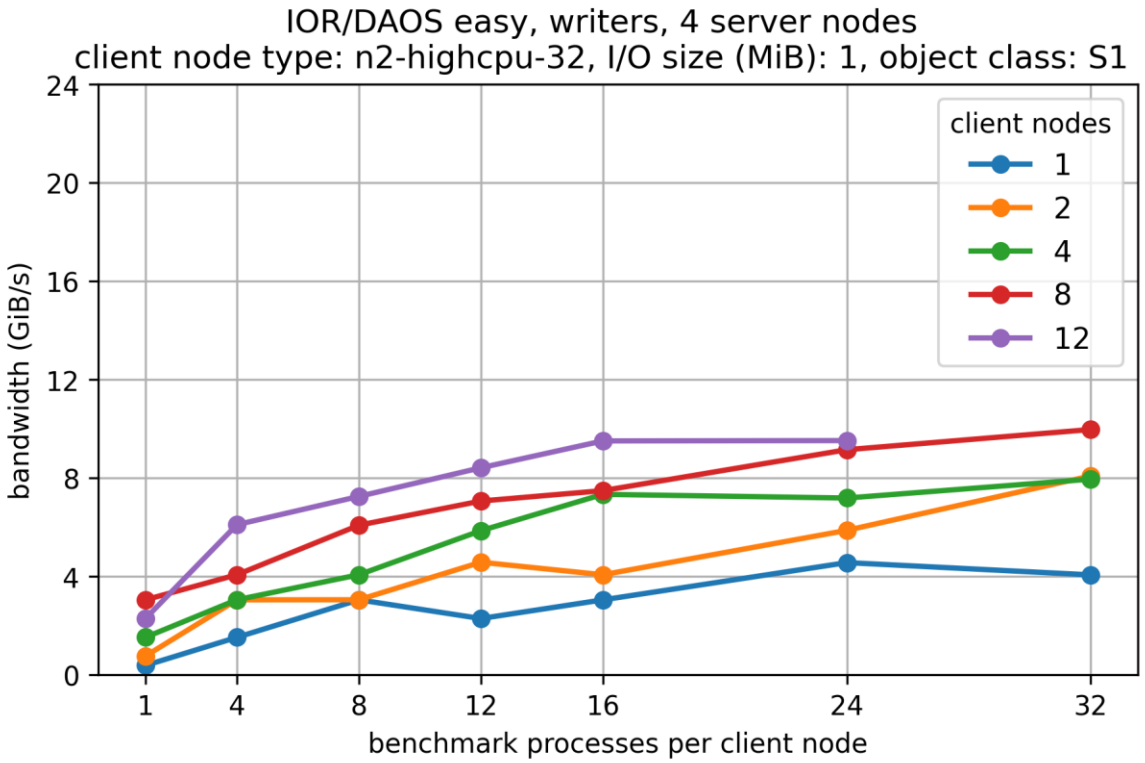
- c3 client instances each with 8 vcpus and a 2.5 GiB/s NIC
- Object class: S1

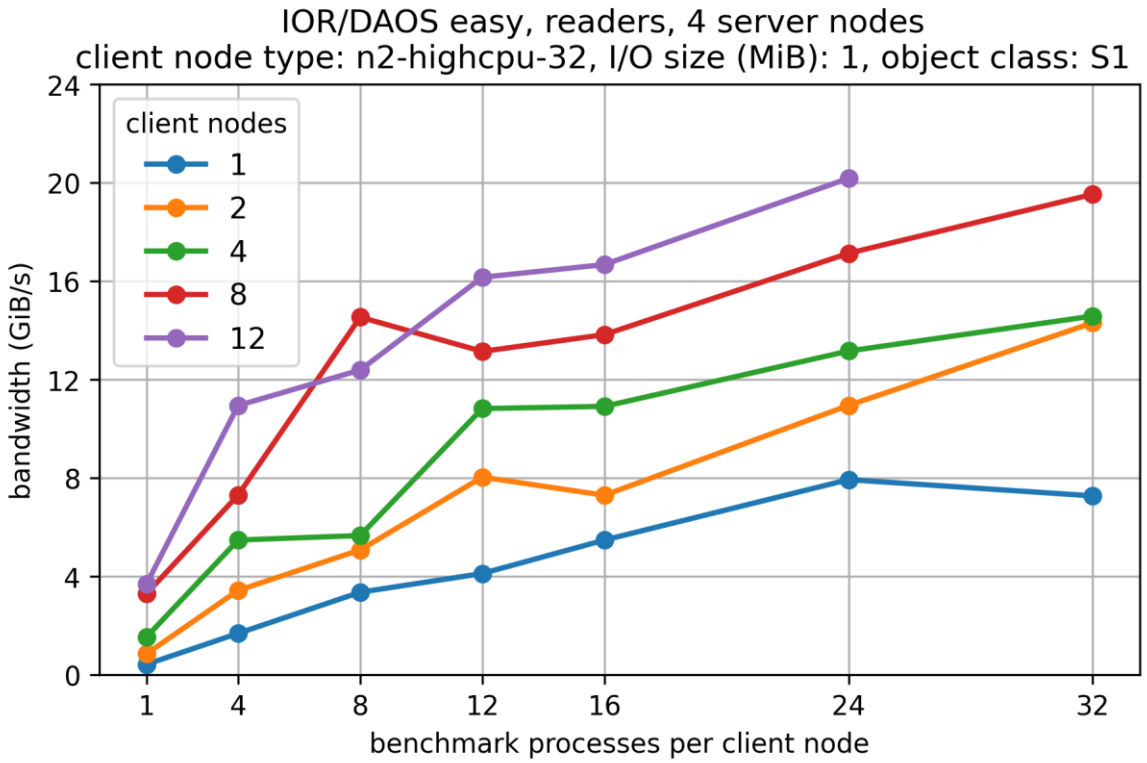




IOR/DAOS on **n2** instances

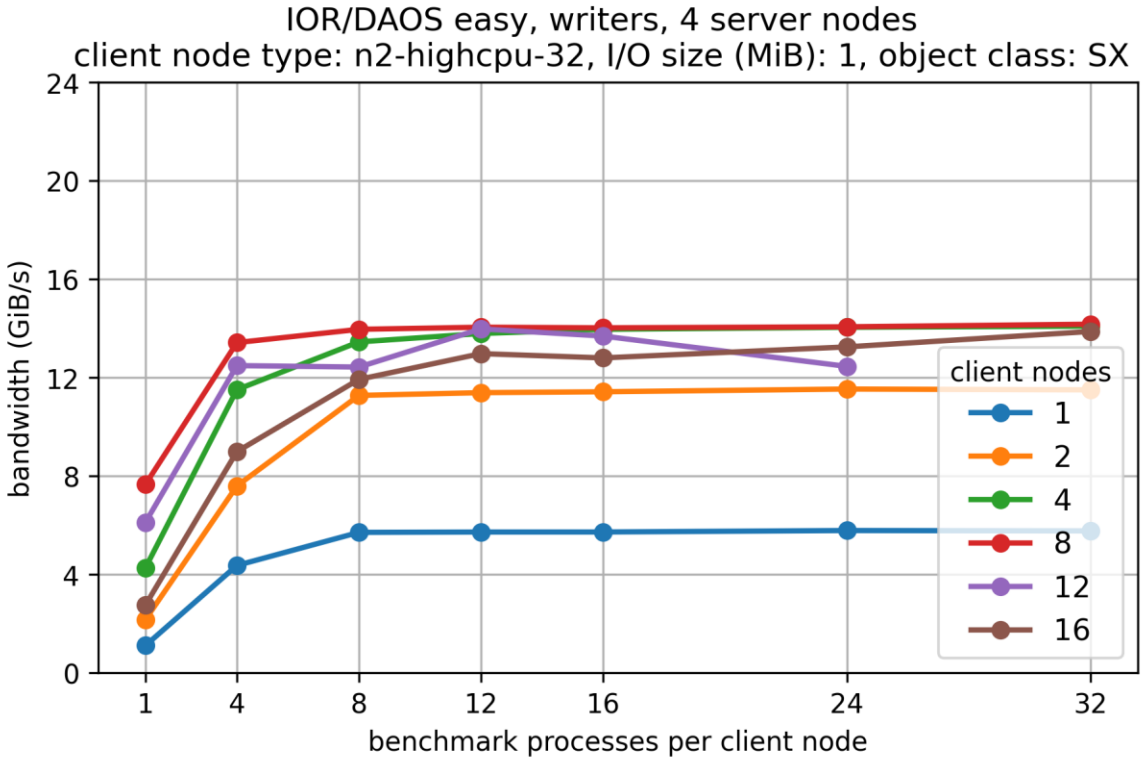
- n2 client instances each with 32 vcpus and a 6 GiB/s NIC
- Object class: S1

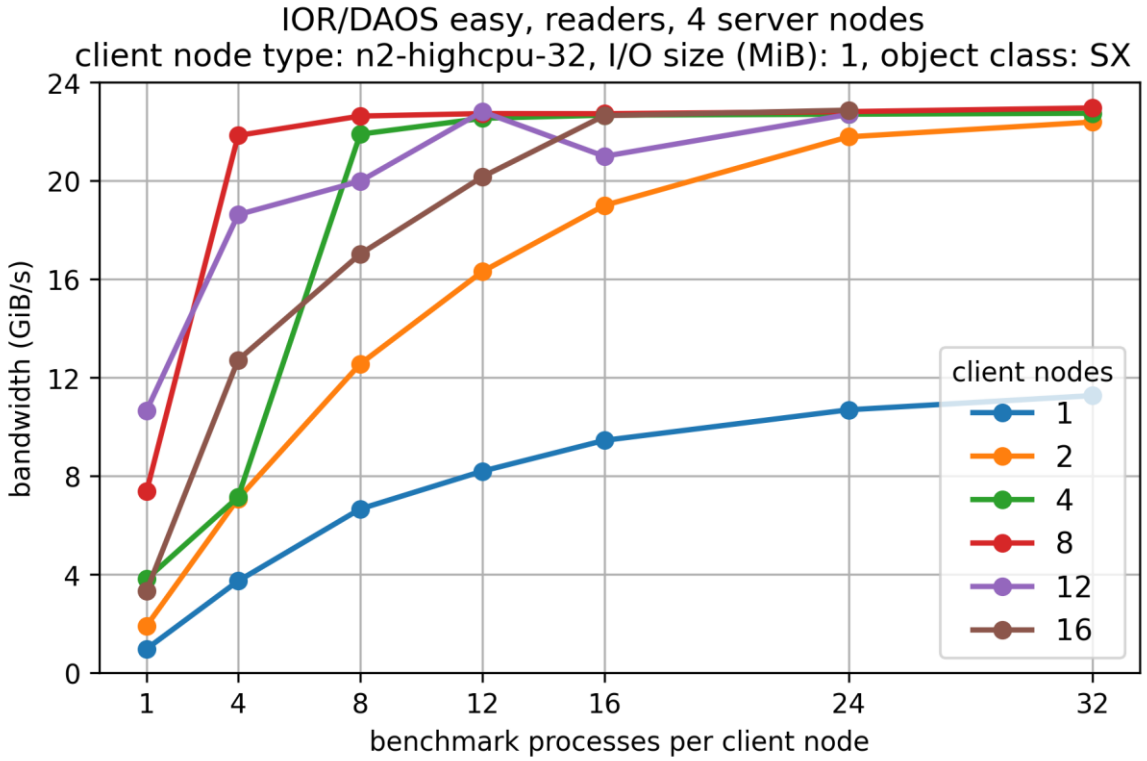




IOR/DAOS on n2 instances

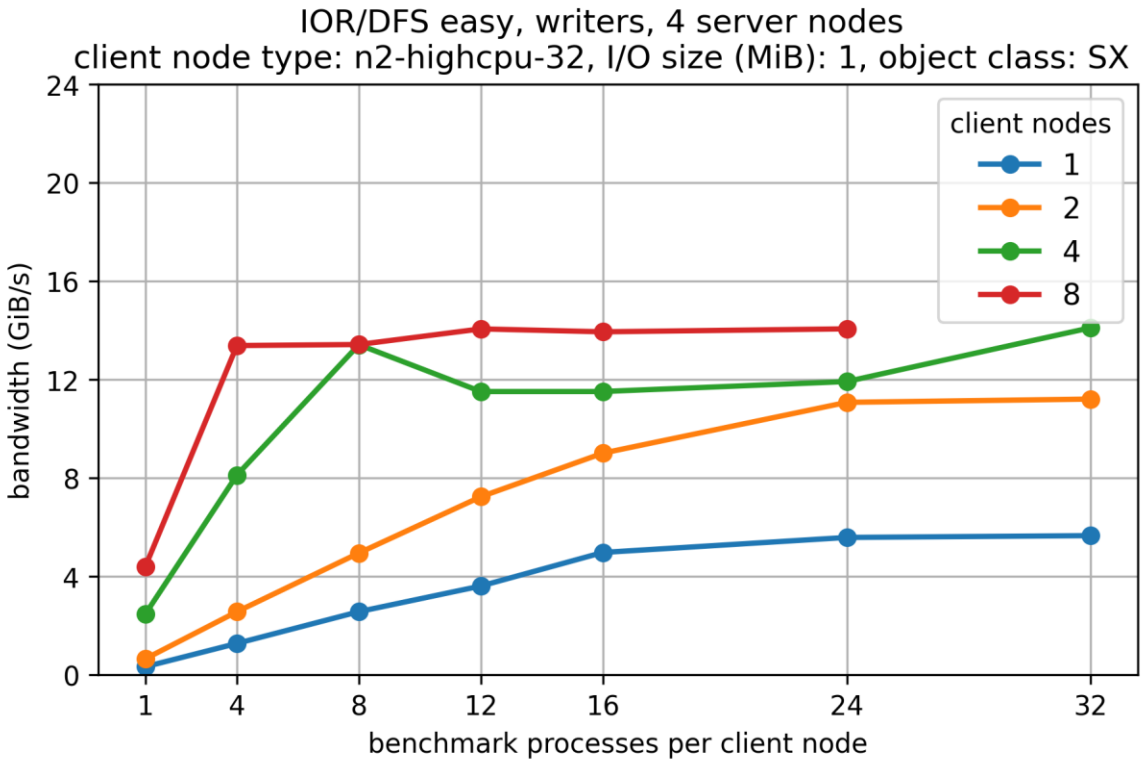
- n2 client instances each with 32 vcpus and a 6 GiB/s NIC
- Object class: **SX**

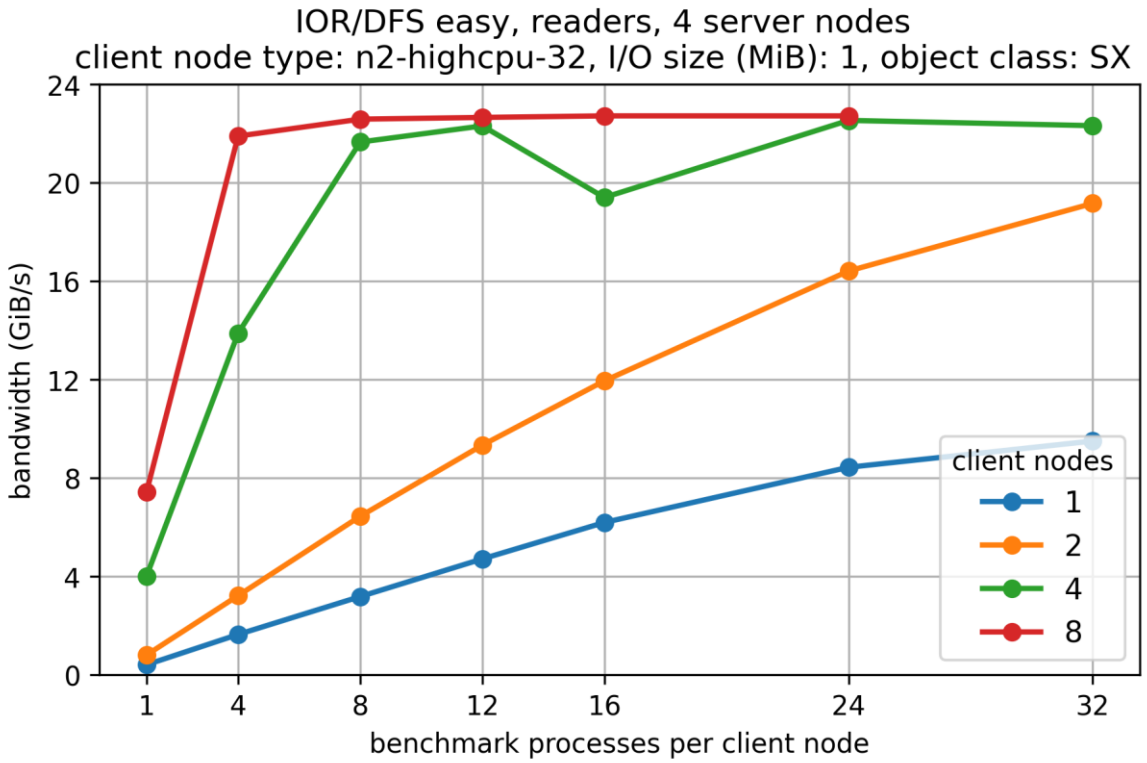




IOR/**DFS** on n2 instances

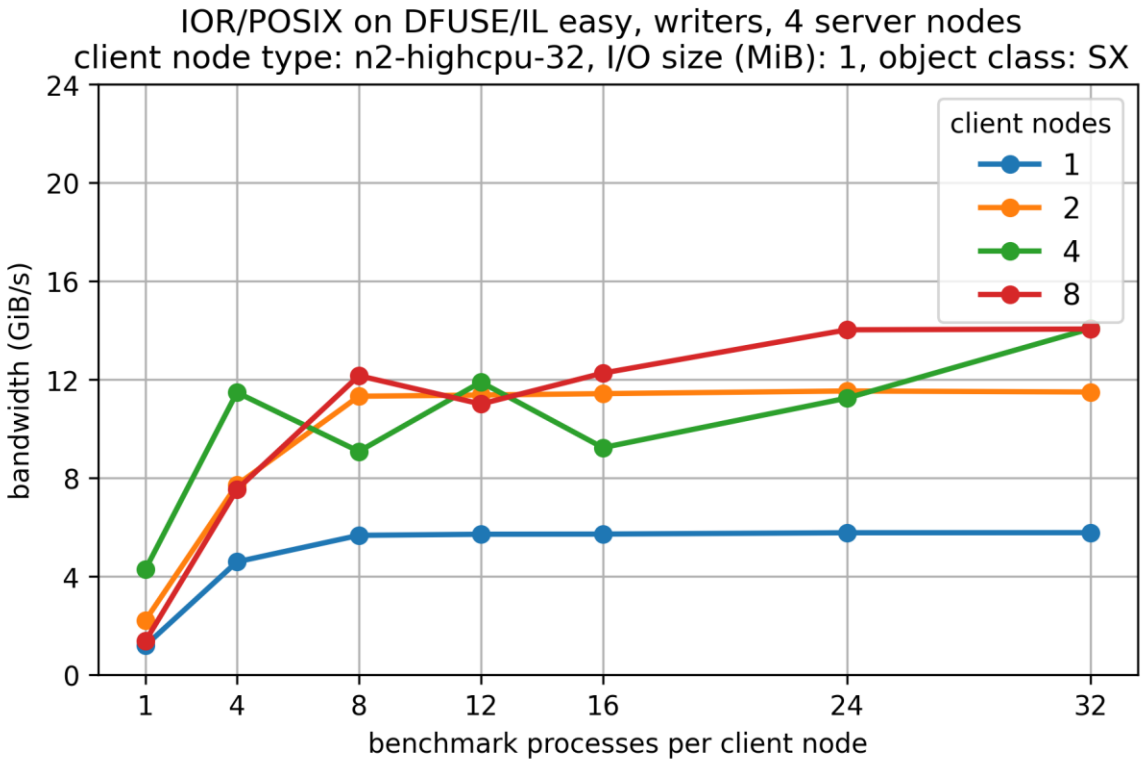
- n2 client instances each with 32 vcpus and a 6 GiB/s NIC
- Object class: SX

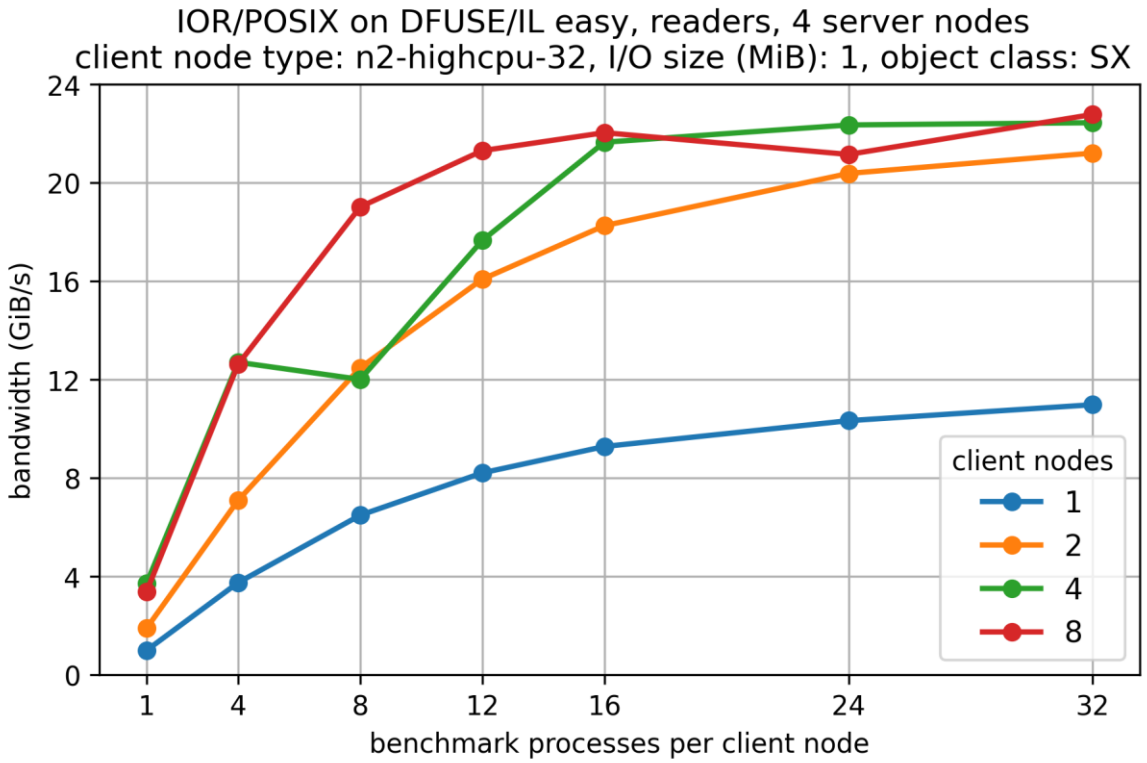




IOR/**POSIX on DFUSE+IL** on n2 instances

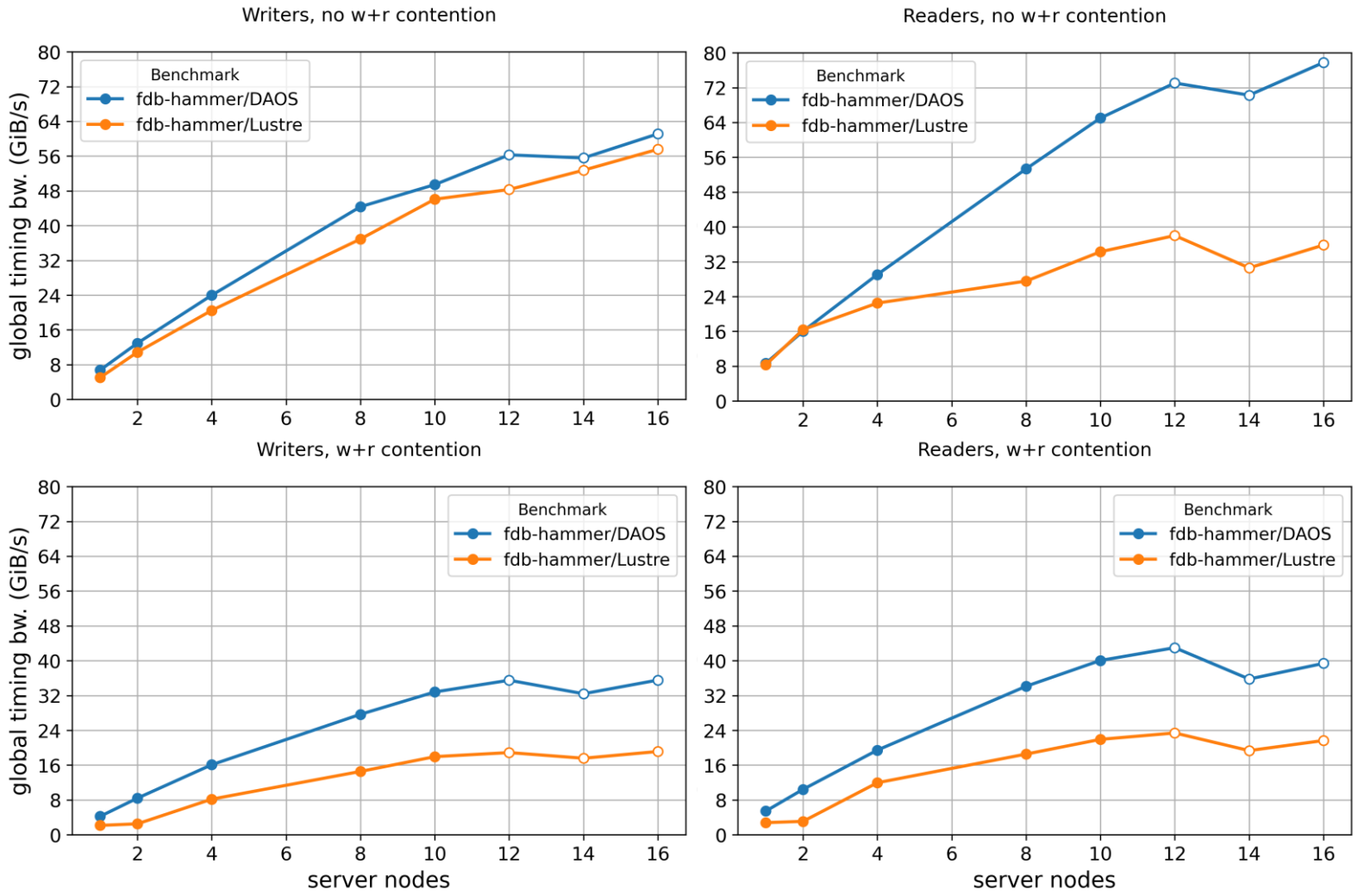
- n2 client instances each with 32 vcpus and a 6 GiB/s NIC
- Object class: SX





PASC'24: ECMWF's DAOS backend vs. POSIX

- Not GCP or ParallelStore
- Complex implement
  - Multiple containers
  - Many keys
  - Range of data sizes



To be presented at PASC'24, 3-5 June.
Paper available at http://www.arxiv.org/abs/2404.03107